School of Informatics

Informatics Research Review
Using NLP to Study the Ancient Egyptian Language

Ricardo Muñoz Sánchez
January 2020

Abstract

The study of Ancient Egyptian language can give us important insights into history, linguistics and the evolution of technology, theology, science and literature through time. Because of this, the use of NLP techniques could be useful to help us extract more information faster than we already do. In this literature review, we will present a survey of the current state of the art on the use of NLP methods to study the Ancient Egyptian language.

Date: Thursday 23rd January, 2020

Supervisor: Maria Wolters

1 Introduction

The Ancient Egyptian culture has been often called the cradle of western civilization. However, there is still much that we do not know about it. The Egyptian people left behind vast amounts of primary textual sources, which the dry weather of the desert helped preserve. As an example of this, we can take the Oxyrhynchus papyri, a collection of over 500,000 papyri that containing fragments of texts, currently housed at the University of Oxford.[1] All of these documents can give us invaluable insights into the lifestyles that these people led and the state of the world at that time. It also can provide unique insights into the how technology, science and religion have evolved over time.

However, some notorious issues are quick to appear. First and foremost is the problem that there are no longer any native speakers left. This means that we cannot know how the language was pronounced or clarify any doubts we may have about the documents. As for making linguistic annotations and translations, it will often take much longer than for living languages. Furthermore, some of the subtleties of the text might be missed.

Another major issue is that the Ancient Egyptian language was used for over 3,000 years. Even worse, the vast expanse of the Ancient Egyptian empire and the lack of quick and inexpensive mediums of transportation lead to major variations in the language. More details on the language and on these variations will be discussed in section 2.1.

Finally, even though a lot of documents survived, most of them are at least partly damaged due to weather conditions, human intervention or just the passage of time. This means that, even if we can extract the whole meaning of the sentence, some nuances or regional variations can be lost to history.

All of these issues mean that the different variations of Ancient Egyptian are considered low resource languages [1, 2]. This means that most of the cutting-edge methods cannot be used for these languages, as those often require vast amounts of data.

For this literature review, we will make a survey of the Natural Language Processing (NLP) techniques that have been used recently to study the Ancient Egyptian language. This includes not only the actual implementations, but also some of the difficulties they faced, how they were able to overcome them and some of the implications of their works.

We will be focusing mainly on Middle and Late Egyptian, but will also devote subsection 2.6 to Coptic, as this language can also be considered a variation of Ancient Egyptian [3] and a good amount of work has been done for it. However, we will not focus on Old Egyptian or Demotic, as practically no NLP work has been done on them. Most of the papers on this literature review are cited not only by other NLP papers, but also by some linguistics ones. We will not talk about the linguistics papers, as most of the papers we will talk about have at least a brief introduction to the language and the relevant properties. We will not talk about optical character recognition or the digital representation of the characters either, as those are image recognition and data representation issues, respectively.

2 Literature Review

For the organization of our literature review, we will first describe the language and make some comments about it in order to showcase common issues that arise when working with

[1] https://www.ees.ac.uk/papyri

the language. We will also talk about the corpora available, including the kinds of annotations they have and the periods over which they have been updated. Then we will talk about the NLP taks that are relevant for Ancient Egyptian. These tasks are automated transliteration, text classification and text retrieval. For each of these, we will mention both the most recent approach to tackle the task, along with any other approaches attempted, when available. It is important to note the lack of part-of-speech tagging and syntactical analysis from this list. This is due mostly to the fact that there has been practically no research done in these tasks [4]. Finally, we devote a whole section to talk about the current state of the use of NLP techniques for Coptic. This is because, even though it is still can be considered an evolution of the Ancient Egyptian language, it has a completely different writing system and we have a greater amount of well-preserved documents. As a result, the issues faced when dealing with Coptic are different than those that we face with Ancient Egyptian.

2.1 The Language

Nederhof and Rahman [2] give a good overview of the Ancient Egyptian language and its characteristics in their paper and is the main source of the information for this section, along with [5]. However, most of the papers that we mention throughout this literature review also have a brief explanation of how the language works.

Ancient Egyptian is a language in the Afro-Asiatic family. This family includes the Semitic languages (Hebrew, Arabic, etc.). In the languages of this family, the vowels are usually not written and Ancient Egyptian is not different. This, coupled with the fact that there are no native speakers alive, means that we cannot really know how Ancient Egyptian sounded. Some of the approximations we currently have are made taking into account how phonetics work in the other languages of the family, but we should not fall into the trap of considering them how the language actually sounded.

The writing system was hierogliphic, but it could also be written in hieratic, a manuscript version of hierogyphs. The symbols of this writing system can be divided into logographs, phonographs, determinatives or typographical signs.

Logographs represent either whole words or ideas. That means that a single symbol can represent a river or a bird. Phonographs, on the other hand, represent sounds. Each symbol corresponds from one to three consonants. Determinatives help clarify the meaning of the word or disambiguate between otherwise identically written words. Finally, typographical signs are used to give semantic meaning to the word or as fillers.

Some important issues arise when trying to parse these symbols. As in Japanese, some words can be written either using logograms, just phonograms or a combination of the two. Also, some symbols can have more than one function and there are neither end-of-word nor end-of-sentence markers. Furthermore, scribes took into account the aesthetic value of their work, adding or removing symbols as they deemed appropriate. Along the same vein, while the language was written from top to bottom, it could be written from left to right or from right to left and the orientation of the text could be either vertical or horizontal. This means that there is no standardized way of writing the language.

The language also had important variations throughout its history. The Ancient Egypt empire lasted for around 3,000 years and is usually divided into the Old, Middle and New Kingdoms. Between these kingdoms there were periods of great unrest, which lead to big cultural changes. Because of that, the Ancient Egyptian language can be divided into these same stages, with Old and Middle Egyptian being sometimes grouped into Classical Egyptian due to their similarity. However, Late Egyptian does show important differences with Middle Egyptian, both grammatical and morphological, and is often considered as a different language.

Finally, Demotic and Coptic can also be considered later stages of Ancient Egyptian, even though they no longer use neither hieroglyphs nor the hieratic script [3]. They can also have bigger variations in terms of morphological and grammatical variation, as evidenced by the greater amount of usage of suffixes and the lack of repetition of phonemes in Coptic [1].

It is because of all these reasons that most papers just focus on one of the stages of the language instead of trying to apply it to all of its history.

2.2 Corpora

An important first step in order to being able to use Natural Language Processing is to have annotated corpora. However, when studying ancient languages we have the major issue that there are no longer any native speakers to annotate sentences or documents. This in turn means that it takes much longer for them to be annotated. When we take into account the fact that Ancient Egyptian is a low resource language, we also face the issue that automated annotations become much less reliable. Here we present the most recent and most comprehensive corpora for the different stages of Ancient Egyptian that we mentioned in the previous section.

The Ramses project is the most ambitious project regarding Ancient Egyptian corpora, as it is an attempt to build a comprehensive annotated corpus of all available texts in Late Egyptian (c. 1350-700 BC). The project began in 2008, and a first version of their software was first made publicly available in 2013 by Polis et. al. [6]. A beta of an online version was released in 2015 [7]. At the time of its presentation, the corpus had already more than 1350 texts, which amount to over a million words. However, when the website was announced, it already had over 4000 texts and, during a presentation in 2017 [8], it was nearing 5000 texts.

An important feature of this corpus is that from its inception, it included the documents that are considered the most useful for studying the language, along with other texts considered to be important for linguistic analysis. The corpus's annotations focus heavily on inflections, lemmata, and spellings, but it also includes all of the relevant metadata to each text, along with annotations on the state of preservation of the documents (or sections of them) and on alterations or editings of the texts. It also allowed the annotators to include comments or criticism on their choices, with references that justify their choices. Their original paper also includes a small tutorial on how to use their software and a list of ways to further expand the project, one of which was including syntactic analysis of the texts.

The online version is currently available at the project's website.[2] However, this is only the beta version of the website, which is only availabe in French and provides access to only a small portion of the corpus. This means that not all people will be able to use the whole corpus both due to the language barriers and due to lack of free access to it. Another major issue is that the las update to the website was made in 2016, though Polis and Razanajao [8] noted in 2017 that the project was still alive.

While there were attempts at making corpora of annotated Middle Egyptian, it was until 2017 that Nederhof and Rahman [2] annotated a corpus for hieratic transliteration that also included the function of each symbol. Taking into consideration that the current models do not use the spatial relations of the script, they linealized the text. They also removed variations of symbols, considering that they would do more harm than to help training the models. The

[2] http://ramses.ulg.ac.be/

corpus currently consists of only two texts. Due to how some words tend to be often repeated throughout each text, its creators suggest to train it on one of them and test it in the other. They argue that mixing both texts would most likely skew the results and give a false sense of confidence. The corpus is available as part of the larger St. Andrews corpora.[3]

The Thesaurus Linguae Aegyptiae [9] was a corpus released in 2004 and was updated until 2012. It contains a wide variety of texts, ranging all the way from the Old Kingdom to the Roman times, including the oldest pyramid texts. This amounts to almost a million and a half words, containing texts in Old, Middle and Late Egyptian and Demotic. It is one of the few annotated Old Egyptian and Demotic corpora. However, the corpus only has lemmatization and morpho-syntactic annotation and most of their website, including the handbook on how to access and use the database, are in German. This greatly limits the amount of people that can access it and the uses that it can be put to. The corpus is freely available online.[4]

The Chicago Demotic Dictionary [10] is another of the few corpora available for Demotic. It was maintained and updated from 1972 to 2012 and includes not only the words themselves, but also scans of the actual documents. The 2002 edition can be found on the project's website as a PDF document.[5]

Finally, a comprehensive corpus of Coptic, was created in 2013 and released in 2016. This corpus, called the Coptic Scriptorium [11], was designed to be used to study a wide variety of subjects, from linguistics to biblical studies, and consists on eleven smaller corpora. At the time of its release, it had a little less than 60 thousand manually annotated words. This corpus can be used for a wide variety of NLP tasks, most of which can be consulted at the project's website.[6] Most notably, it covers a wide variety of annotations, from tokenization (i.e. identifying the words in a document) all the way to parts-of-speech tagging and a treebank which follows the universal dependencies notation. This is an ongoing project that currently has around 850 thousand annotated words and the documents have enough metadata to tell whether these annotations were made automatically or whether they were either made or revised by humans. Their most recent release was on September 2019 and the current status of the project can be found at their blog.[7]

2.3 Transliteration

We currently have a very good understanding of how Ancient Egyptian script works, even going as far as having developed standardized methods of transliteration and designed Unicode symbols for hieroglyphic script. However, most of these transliteration methods require human annotators to work on the text because of the lack of standardization of the language mentioned in section 2.1. This means that transliteration is still an open problem in the Ancient Egyptian machine learning field.

An important issue that arises when using human annotators is that it is a slow process, which becomes even more slower due to there being no more native speakers of the language. Because of this, any major breakthrough would mean that more manpower would be available for other tasks in Egyptology.

Nederhof is a researcher that has been focusing in this area for several years. His most recent

[3] https://mjn.host.cs.st-andrews.ac.uk/egyptian/texts/

[4] http://aaew.bbaw.de/tla/

[5] https://oi.uchicago.edu/research/projects/chicago-demotic-dictionary-cdd-0

[6] https://copticscriptorium.org/tools

[7] https://blog.copticscriptorium.org/

paper is with Rahman, from 2017 [2]. They made a probabilistic automaton that can transcribe a text in Middle Egyptian hieratic (i.e. manuscript hieroglyphs) to its phonetic values. For this, they created the Middle Egyptian mentioned in section 2.2. It has annotations for the functions of each symbol that appears so as to help the model learn. They consider that the innovation of their system is that does more than just doing a simple transliteration, it also makes notes on semantic elements of the text. Due to the scarcity of annotated texts from that era, they had to build an n-gram model (i.e. a model that only takes into account the previous n characters) and were able to reach recall and precision scores of approximately 0.95. The authors mention that, even though the model used was very basic, this is an important steppingstone for transliterating documents from this era.

One of the issues that with this paper is that its author's objective was not completely clear just by reading the abstract. They also do not make any mention of the corpus that they had to create for it until halfway through the paper. On the positive side, they give a very thorough introduction to the structure of language and explain their model well enough so that it could be reproduced. They also give references for several frameworks that attempt to do the same as them, but in different languages (for example, when the exact same thing can have several written forms), and of different approaches that have been used and how those helped shape their choices in this paper. As for understanding the paper, while most of the decisions of the model and how it was implemented were clear, we feel that a deeper understanding of the language would be necessary to be able to do any further work than the examples provided in the paper, especially when evaluating the correctness of the transliteration.

Rosmorduc [12] tried another approach to transliteration. He derived a set of rules on how words are formed and created a series of transducers, that is, finite-state automatons that parse the words and use these rules to verify whether a word is valid or not. The validation set was one of the same texts that Nederhof and Rahman used for their corpus and his model achieved a precision of around 0.91. However, this was the same set from which the rules were derived. When using another text as a test set, the precision drastically dropped all the way to 0.82. However, he justifies his results by claiming that they were due to some small technical errors. Finally, he tried to use the same model on a Late Egyptian text. Even though the precision score for this test is not reported and the author notes that it is quite bad, he mentions that it is on par on what he would expect for a student that has only studied Middle Egyptian but not any of its latter variants.

Even though the author was very clear on is purpose from the very beginning, the paper still has some glaring issues. Most of the paper is spent explaining the rules that will be fed to the transducer, but neither the notation nor the typography help make the actual contents easier to read. Maybe some diagrams could have helped make the paper more readable, but some of the rules might have been a bit more complicated than what could be easily illustrated with diagrams. Another major issue was that there was no actual reporting on the performance on the Late Egyptian text. There was no precission score nor any sort of data or comparison to justify the claims that the results were on par with what would be expected from a student.

A later paper by Barthéley and Rosmorduc [13] compares two kinds of transducers, but no there is neither a performance given for either model nor an actual comparison made between the two. Even though the authors justify that as some bad rule choices in the transducer's rules, that provides no useful insight for a reader of the paper.

2.4 Text Classification

Automatic text classification is another important task in NLP, as it can help document organization and management, text filtering or sense disambiguation. Gohy et al. [14] also claim that doing text classification can also give us insights into the registers used for different kinds of texts, which in turn should help improve the performance of machine learning techniques in other NLP tasks. They further claim that this is a more important endeavor in the case of dead languages such as Late Egyptian.

More specifically they did genre classification in their paper. This is a special case of text classification that assumes that the genres in which the texts will be classified are assumed to be mutually exclusive from each other Another characteristic of genre classification if that the tags are aleady given to the learning algorithm. The authors do note this is often an over-simplification, however they also mention that, when chosen carefully, the genres should be relatively independent from each other. The genres chosen for their paper were letters, judicial documents, oracular questions, educational texts, monumental inscriptions, hymns and administrative texts. They also note that another strong assumption that they are making in their paper is that each genre will have one and only one register and that each register will be exclusive to one genre, which is not true in general. Finally, as they are only interested in the registers, their models use mainly just semantic and morpho-syntactic features, while mostly ignoring the metadata and the structure of the texts.

The models that they used were a naïve Bayes classifier (which assumes that each feature is independent), a support vector machine (which maps the documents to a vector space and then classifies the documents depending on their relation in that space), and a segment and combine method (which learns from each syntactic property of the document and then combines what it learnt to get further insights). Their best performing model was the naïve Bayes classifier, which achieves a recall of slightly over 0.84 in general and of over 0.97 with both letters and monumental inscriptions. They consider that in the case of the monumental inscriptions this is due to the more rigid structure used for the language and in the case of the letters it is due to the higher volume of training data available. On the other hand, this model gets a recall of only 0.66 with oracular texts. The authors consider that this is because oracular questions were usually very short. Therefore, they created a modified naïve Bayes classifier which takes into account the length of the texts. This new model improved the recall of oracular questions to over 0.9 and got a general recall improvement of approximately 3%. Their support vector machine model got similar, but slightly worse results than the naïve Bayes classifiers and the segment and combine model got much more extreme results, with letters, judicial and educational documents, and monumental inscriptions getting a recall of over 0.9, but oracular questions and administrative texts having a recall lower than 0.3.

In general, I think that their paper was well organized and their results were well presented. Most of the relevant concepts were briefly explained so that someone that was not an expert on the field could still understand what they were talking about without clogging the paper. They also analyzed the different models that were proposed and gave possible explanations for the behaviors observed. However, they never set a baseline model and throughout the paper they keep talking about the "performance of the model" without further explaining in which sense they meant it. Finally, even though they claim that their work could be used to improve the performance of other NLP tasks, no clear examples or insights on how that could be done are given. This especially problematic considering that Ancient Egyptian is a low resource language. They do not talk much about any further work that could be done on this task, only mentioning that their segment and combine model was promising.

2.5 Text Retrieval

One of the NLP tasks that would be the most useful for egyptologsts is text retrieval. This task allows to create systems capable of searching and querying indexed documents. Using these kinds of systems would save researchers the effort of sifting through piles of useless data. They also function as a cultural preservation tool, by diminishing the amount of manipulation suffered by each individual document.

In their paper, Iglesias-Franjo and Vilares [15] created a text information retrieval system for Middle Egyptian. For their work, the authors consulted several egyptologists in order to determine the needs of such a system. Most of these needs are either simplicity of use, flexibility and adhering to the current standard practices of the field. The system first preprocesses and normalizes the text of the documents. After this, an index is created and stored. Once the index is in place, queries can be made. These can be made in latin script, hieroglyphs or a combination of the two. The text is then normalized as in the indexing stage, with the difference that a query using hieroglyphs can specify whether the symbols are the only ones appearing or if the user is looking for a words that contain those symbols. Then, a list is selected and ranked according to a Boolean model and a vector space representation of the documents. The authors note that this is a first release and that there is still much work to be done. The system is freely available at their GitHub page.[8]

The authors clearly explain their design choices throughout the paper, making special emphasis on who the target audience of the system is (the egyptologists). They also give a quick but clear overview of the language and how it has been historically typewritten into computers without it taking too much space. One of the few criticisms of their paper would be that, even though the general method for selecting and ranking the documents is mentioned, no further details are given. They do mention some related works, but do not specify if any of them uses the same querying algorithm. Another approach that they proposed was using a method similar to those used for Japanese dictionaries, where words ca be searched for using a combination of kanji (ideaograms) and kana (silabary). However, this query method was considered too unintuitive by the authors. They also note that completion of the Ramses or the Thesaurus Linguae Aegyptiae corpora mentioned in 2.2 could be a great boon to these kinds of systems.

2.6 Coptic

Even though Coptic can be considered a later stage of Ancient Egyptian, it has important differences with respect to Classical and Late Egyptian. This leads to a completely different set of problems when using NLP techniques with the language. One of these differences is that Coptic is no longer written in hieroglyphs, it uses a modified version of the Greek alphabet instead. This means that transliteration is no longer an issue, as there is a one-to-one correspondence between symbols and phonemes. Other factors include changes in the morphology of the language, both in general and more specifically in its written form.

Another important factor is that a lot of documents from early Christianity were written in Coptic and that the Coptic Orthodox Church still uses the language during mass. This means that there are more well-preserved texts in Coptic than in Ancient Egyptian and that the contents of these texts tend to attract more attention from a wider variety of scholars.

One attempt at doing morphological analysis comes from a paper from by Smith and Hulden [16]. They focu only on a dialect of Coptic. They consider that, as it is mainly a prefixing

[8] http://github.com/estibalizifranjo/hieroglyphs

language save for a few notable exceptions, a good model could be a transducer, that is, a finite-state automaton that parses words and determines whether they are correctly formed given the rules of the language. Their testing set was composed of over a hundred words and had a recall slightly lower than 0.95. They think that their work could be useful for teaching the Coptic grammar and mention that it could help study the larger coptic texts. However, they make no mention on whether their model would need major modifications to consider other dialects, only stating that increasing the coverage of their analyser would need more lexicographical work. An important positive aspect of their paper is that, even though the authors don't state the rules for the transducer in it, they do give references on where to find them. Another positive aspect is that the notation used throughout the paper is very clear, which makes it much more easier to understand.

As was mentioned in section 2.2, the Coptic Scriptorium is a corpus that had at its release a little less than 60 thousand words available. Throughout the years, a wide variety of tools readily available for use have been developed for it. These go all the way from a lemmatizer to part-of-speech taggers. As this tools are pretty recent and represent significant advancements in how NLP is used for Coptic, we will talk about the two papers that introduced them first.

Zeldes and Abrams [17] consider that the creation of a treebank compatible with the universal dependency annotation scheme would be an important addition to the study of Coptic in general. They decided to work with the Coptic Scritptorium corpus due to it being freely available and that the automatic segmentation achieves a very high precision score, which means that it can be considered a gold standard. They mainly decided to follow two main principles: when possible their notation should compatible with the previous literature in the field and they would try to keep the notation in line with the practices in Hebrew and Arabic, which come from the same language family. When testing their treebank against expert human annotators, they got an agreement of over 95%. The agreement dropped to slightly over 0.85 when compared to undergraduate students.

A good thing about this paper it that it explains the choices they made when building the treebank and notes the possible benefits of having built it. It also mentions that this was the first treebank built for the Egyptian language subfamily and notes that their work should be useful when building ones for other languages of this subfamily, such as the one that the Ramses project aims to have in future stages.

Even more tools for the Coptic Scriptorium came in the form of a complete pipeline for NLP analysis. Zeldes and Schroeder [1] created an online tool that automatizes a lot of tasks, from word and morpheme segmentation all the way to dependency parsing. In their paper, the authors specify how they chose each of their models according to the needs of the tasks at hand. These tasks are segmentation, normalization, tagging and lemmatization, detection of language of origin, and parsing.

Segmentation is when a word is separated into its lexical sub-units. This is especially clear in languages from the Afro-Asiatic family, as they tend to be somewhat agglutinative. For this task they selected around 180 segmentation rules and created a model that determined the priority order of the rules through 10-fold cross-validation. The accuracy of this model was slightly higher than 0.9. In the normalization stage, they had to consider the use of diacritics, spelling variations, and abbreviations. For this task, they used a combination of a predetermined list of common variations and a learnt list of the use of diacritics and capitalization. This model had an accuracy of 0.98. For parts-of-speech tagging and lemmatization, used an algorithm called TreeTagger [18] and achieved accuracies of 0.95 and 0.97, respectively. As for determining the language of origin, they had an accuracy of over 0.93. Finally, the parsing section has a preliminary version of the model of the paper from Zeldes and Abrams mentioned previously in this section and achieves an accuracy of 0.87.

Each of the components on this paper can be used either on its own or as part of a pipeline and can be accessed both at the author's website[9] or as part of the Coptic Scriptorium project.

3 Summary & Conclusion

As we have seen, the use of NLP methods on Ancient Egyptian is important both from linguistic and historical reasons. However, the advances in this area have been sparse though time. This is in part due to it being a dead language (which means that it takes much longer to translate and to annotate than other languages would) and in general to it being a low resource language (which means that the state of the art neural networks cannot be easily used, if at all).

Rosmorduc [4] gives a quick overview of some of the main tasks that have been tackled from the 90s to 2015, including the task of representing the language on computers. He notes that, other than some attempts in the 90s, most of the work up until recently had been geared towards creating a standard Unicode representation of hieroglyophs. The most recent update on this regard was on 2019 [19], when some control characters to signal some spatial properties of the characters were introduced.

Polis [6] and Nederhof and Rahman [2] consider that this lack of advances has been in good part due to the lack of annotated text, but also note that most attempts are trying to generalize over large periods of time, even when taking into account divisions such as Middle and Old Egyptian.

Another one of the major issues is that throughout time, most papers have focused on Coptic. This is understandable, as the texts in this language are much better conserved and it is still being used during masses of the Coptic Orthidox Church, albeit in a limited capability. However, this tends to shift attention from the other stages of Ancient Egyptian, with Demotic being the most affected.

In his 2017 talk, Polis [8] also notes that more interaction between projects could be useful, not only in the field of computational linguistics, but in Egyptology as a whole. This is especially important as most projects end up using either the same datasets or the same objects, but end up having their own systems that are not compatible with each other. An example he gives is that of an statue with inscriptions. The artifact itself has value for some researchers, while the kind of object might be of interest to others. He also notes that, while some researchers might be interested in the location and the layout of the text, some others might be just interested in the text itself or even in just the content. He notes that there is a current collaborative project called THOT [20] that is looking to be a bridge for these areas of study. However, the project does not have any sort of connection to the actual databases as of 2019 and does not have any kind of roadmap to show how it will grow in the future.

Another thing of note is that this area of research appears to be approached by a very limited amount of researchers. This might indicate a lack of funding or of interest in this area. However, some of these research groups appear to be growing, such as the one dedicated to the Ramses corpus, the evolution of which can be seen in [6], [7], and [8]

As a final note, an interesting thing would be to compare and contrast the NLP advances that have been done in other ancient languages, such as Sumerian, Ancient Greek, Sanskrit,

[9] https://corpling.uis.georgetown.edu/coptic-nlp/

etc. This could show how the advances in done in these different languages have affected or influenced each other. Even though some of the papers that we have mentioned so far did show this awareness, most did not. An NLP package called The Classical Language Toolkit [21] is being developed. It has tools for several ancient languages and even provides access to corpora for several of them, including the Coptic Scriptorium corpora mentioned in section 2.2. This package should help encourage more research on these languages, which will help in turn gain important insights into our past.

References

[1] Amir Zeldes and Caroline T. Schroeder. An NLP Pipeline for Coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 146–155. Association for Computational Linguistics, August 2016.

[2] Mark-Jan Nederhof and Fahrurrozi Rahman. A Probabilistic Model of Ancient Egyptian Writing. In *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing 2015 (FSMNLP 2015 Düsseldorf)*. Association for Computational Linguistics, 2015.

[3] Kathryn A. Bard. Egyptian language and writing. In *Encyclopedia of the Archaeology of Ancient Egypt*, pages 325–328. Routledge, November 2005.

[4] Serge Rosmorduc. Computational Linguistics in Egyptology. *UCLA Encyclopedia of Egyptology*, April 2015.

[5] Janice Kamrin. *Ancient Egyptian Hieroglyphs: A Practical Guide - A Step-by-Step Approach to Learning Ancient Egyptian Hieroglyphs*. Harry N. Abrams, November 2004.

[6] Stéphane Polis, Anne-Claude Honnay, and Jean Winand. Building an Annotated Corpus of Late Egyptian. The Ramses Project: Review and Perspectives. In *Texts, languages & information technology in egyptology: selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptolgie), Liège, 6-8 July 2010*, Aegyptiaca Leodiensia, 2013.

[7] Stéphane Polis, Serge Rosmorduc, and Jean Winand. Ramses goes Online. An annotated corpus of Late Egyptian texts in interaction with the Egyptological community. *International Congress of Egyptologists XI*, August 2015.

[8] Stéphane Polis and Vincent Razanajao. Ancient Egyptian philology: The digital turn. Current projects and future perspectives for the study of Ancient Egyptian texts. *International Congress of Egyptologists XI*, February 2017.

[9] Stephan J. Seidlmayer. Handbuch zur Benutzung des Thesaurus Linguae Aegyptiae (TLA), 2011. http://aaew.bbaw.de/tla/.

[10] Janet H Johnson. *The Demotic Dictionary of the Oriental Institute of the University of Chicago*. The Oriental Institute, University of Chicago, 2002.

[11] Caroline T. Schroeder and Amir Zeldes. Coptic SCRIPTORIUM:A Corpus, Tools, and Methods for Corpus Linguistics and Computational Historical Research in Ancient Egypt, December 2016. https://copticscriptorium.org/download/copticscriptorium-HD-51907-14-whitepaper-rev.pdf.

[12] Serge Rosmorduc. Automated Transliteration Of Egyptian Hieroglyphs. In *Information Technology and Egyptology in 2008*. Gorgias Press, December 2009.

[13] François Barthélemy and Serge Rosmorduc. Intersection of Multitape Transducers vs. Cascade of Binary Transducers: The Example of Egyptian Hieroglyphs Transliteration. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pages 74–82, Blois, France, July 2011. Association for Computational Linguistics.

[14] Stéphanie Gohy, Benjamin Martin, and Polis Stéphane. Automated text categorization in a dead language. The detection of genres in Late Egyptian. *Texts, languages & information technology in egyptology: selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptolgie), Liège, 6-8 July 2010*, 2013.

[15] Estíbaliz Iglesias-Franjo and Jesús Vilares. Searching Four-Millenia-Old Digitized Documents: A Text Retrieval System for Egyptologists. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 22–31. Association for Computational Linguistics, August 2016.

[16] Daniel Smith and Mans Hulden. Morphological Analysis of Sahidic Coptic for Automatic Glossing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2584–2588, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[17] Amir Zeldes and Mitchell Abrams. The Coptic Universal Dependency Treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 192–201, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[18] Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*, 1995.

[19] Mark-Jan Nederhof, Stéphane Polis, Serge Rosmorduc, and Simon Schweitzer. Unicode Control Characters for Ancient Egyptian. In *12th International Congress of Egyptologists*, Cairo, November 2019.

[20] Vincent Razanajao. Thot - Thesauri and Ontology for Ancient Egyptian Resources. http://thot.philo.ulg.ac.be.

[21] Kyle P. Johnson, Luke Hollis, and Patrick J. Burns. CLTK: The Classical Language Toolkit, 2014. https://github.com/cltk/cltk.